Optimizing the rapid measurement of detection thresholds in infants

Pete R. Jones
Institute of Ophthalmology, University College London (UCL), London, UK

Sarah Kalwarowsky
Institute of Ophthalmology, University College London (UCL), London, UK

Oliver J. Braddick
Department of Developmental Science, University College London (UCL), London, UK

Janette Atkinson
Department of Experimental Psychology, University of Oxford, Oxford, UK
Department of Developmental Science, University College London (UCL), London, UK

Marko Nardini
Department of Psychology, Durham University, Durham, UK
Institute of Ophthalmology, University College London (UCL), London, UK

Accurate measures of perceptual threshold are difficult to obtain in infants. In a clinical context, the challenges are particularly acute because the methods must yield meaningful results quickly and within a single individual. The present work considers how best to maximize speed, accuracy, and reliability when testing infants behaviorally and suggests some simple principles for improving test efficiency. Monte Carlo simulations, together with empirical (visual acuity) data from 65 infants, are used to demonstrate how psychophysical methods developed with adults can produce misleading results when applied to infants. The statistical properties of an effective clinical infant test are characterized, and based on these, it is shown that (a) a reduced (false-positive) guessing rate can greatly increase test efficiency, (b) the ideal threshold to target is often below 50% correct, and (c) simply taking the max correct response can often provide the best measure of an infant's perceptual sensitivity.

Introduction

Psychophysical tests are often designed to measure an observer's *threshold*—the lowest stimulus level at which the observer performs the task correctly on X% of trials. For example, one might measure the faintest sound an observer can detect 50% of the time or the lowest-contrast image he or she can identify with 75% reliability.

The gold standard measurement of threshold is provided by a *psychometric function*. This maps every stimulus level to expected performance (e.g., percentage correct) and allows bias and inattention to be quantified and compensated for. However, deriving an accurate psychometric function typically requires hundreds of trials, which, with an infant, may entail multiple sessions of protracted testing. Thus, although psychometric functions *have* been used to study early development (e.g., Armstrong, Maurer, Ellemberg, & Lewis, 2011; Banks, Stephens, & Dannemiller, 1982; Dobkins, Lia, & Teller, 1997; Held, Gwiazda, Brill, Mohindra, & Wolfe, 1979; Teller, Mayer, Makous, & Allen, 1982), they are impracticable when perceptual

Citation: Jones, P. R., Kalwarowsky, S., Braddick, O. J., Atkinson, J., & Nardini, M. (2015). Optimizing the rapid measurement of detection thresholds in infants. *Journal of Vision*, *15*(11):2, 1–17, doi:10.1167/15.11.2.

doi: 10.1167/15.11.2 Received December 1, 2014; published August 3, 2015 ISSN 1534-7362

sensitivity must be assessed rapidly, such as in clinics or as part of an extensive test battery.

The pragmatic alternative is to use a rapid, adaptive procedure to estimate one threshold at a single, specified level of performance. For example, a "staircase" rule can be used to adjust the stimulus magnitude up or down until the likelihood of one incorrect answer is equal to the likelihood of two correct answers (70.7% correct; see Leek, 2001; Treutwein, 1995). Adaptive techniques have been widely used in the laboratory to measure a host of abilities in infants, including motion detection and direction sensitivity (Wattam-Bell, 1996), temporal modulation sensitivity (Swanson & Birch, 1990), contrast sensitivity (Banks & Salapatek, 1981), visual acuity (Lewis, Maurer, Chung, Holmes-Shannon, & Van Schaik, 2000), absolute tone detection sensitivity (Werner & Marean, 1991), and auditory masking thresholds (Werner & Bargones, 1991).

However, even adaptive procedures are often too slow or too unreliable for the clinic, in which assessments must be made in a matter of minutes and in which the results must be meaningful even within a single individual. As a result, clinical tests often deviate from their laboratory counterparts in a number of respects. First, numbers of trials are typically reduced. This is done in order to minimize test durations and maximize compliance. Second, the method of computing threshold is often simplified, for example, by defining threshold as the highest or last correct response rather than by averaging over reversals (e.g., Day et al., 2008; McDonald et al., 1985, p. 1158). Simple decision rules are usually justified as making the test more user-friendly for clinicians although we shall argue in the present paper that a simple decision rule may actually make the test more robust also. Third, efforts are often made to increase the amount of information gained from a single trial. For example, the target may be randomly distributed in time (e.g., as in audiograms; Day et al., 2008) or space (e.g., as in a recent test of visual acuity; Jones, Kalwarowsky, Atkinson, Braddick, & Nardini, 2014) or the operator may be allowed to make "a broadly integrative, subjective judgment" (McDonald et al., 1985, p. 1158) regarding whether or not the infant perceived the stimulus (i.e., rather than having to choose between toward which of two locations the infant looked). As discussed in the present manuscript, the implications of such changes are relatively complex. However, in general, they are intended to reduce the likelihood of a response being scored as correct by mistake, making correct answers more informative.

These design modifications can produce tests that function well. For example, the clinical Acuity Card test[1] is around seven times faster than its laboratory equivalent but does not differ significantly in terms of expected scores or test–retest reliability (Mohn & van Hof-van Duin, 1986; Teller, McDonald, Preston, Sebris, & Dobson, 1986). However, the various procedural changes mean that rapid clinical procedures, in contrast with more traditional laboratory measures (Olsho, Koch, Halpin, & Carter, 1987; Teller, 1979), often lack a rigorous grounding in established psychophysical theory. For example, an adult asked to indicate which of two locations contains a stimulus (two-alternative forced-choice [2AFC] detection) will exhibit a false positive ("lucky guess") rate of 50%. Conversely, it is unclear what the corresponding likelihood is that an infant will be judged to have seen a stimulus that appears in one of two locations or how a guess rate that differs markedly from 50% will affect the ultimate reliability of any test results. Furthermore, the difficulties in relating practice to theory are made greater by the particular challenges of working with infants. For example, infants often exhibit high interindividual variability and high levels of inattentiveness (see Banks & Dannemiller, 1987; Swanson & Birch, 1992; Teller, 1979; Viemeister & Schlauch, 1992; Werner & Marean, 1991). In contrast, many psychophysical techniques presuppose a homogeneous cohort of highly attentive observers such that, in adults, a lapse rate greater than 6% is often taken to imply "that the experiment was not performed properly and that the data are invalid" (Wichmann & Hill, 2001, p. 1295).

These differences in both methodology and observer characteristics can cause uncertainty. For example, it is unclear (a) how the statistics of clinical tests differ from those of traditional psychophysical designs (e.g., *m*-alternative forced-choice tasks), (b) what the optimal adaptive method is to ensure fast and reliable convergence on threshold, or (c) whether taking the maximum correct response provides a valid measure of threshold. In other respects, the lack of underlying theory has led to apparent inconsistencies. For example, why is it that clinical adaptive designs often target 33% or 50% correct thresholds (Day et al., 2008; Jones et al., 2014; Widen et al., 2000) when adult observer models consistently recommend targeting thresholds of 83% or above (Green, 1990; Klein, 2001; McKee, Klein, & Teller, 1985)? Finally, in some cases, a lack of theoretical clarity may even have led to undue pessimism with authors suggesting, for example, that rapid behavioral tests are unfeasible in infants younger than 6 months and will invariably "yield unrepeatable results unless the rigors of an observer-based [laboratory] psychophysical procedure are adopted" (Cone-Wesson, 2003, p. 175).

In the present study, we used Monte Carlo simulations, together with empirical data from a recent, automated measure of visual acuity (Jones et al., 2014), to show how the statistics of a clinical infant test can differ from those of traditional (adult) psychophysics and to determine the consequences of these differences

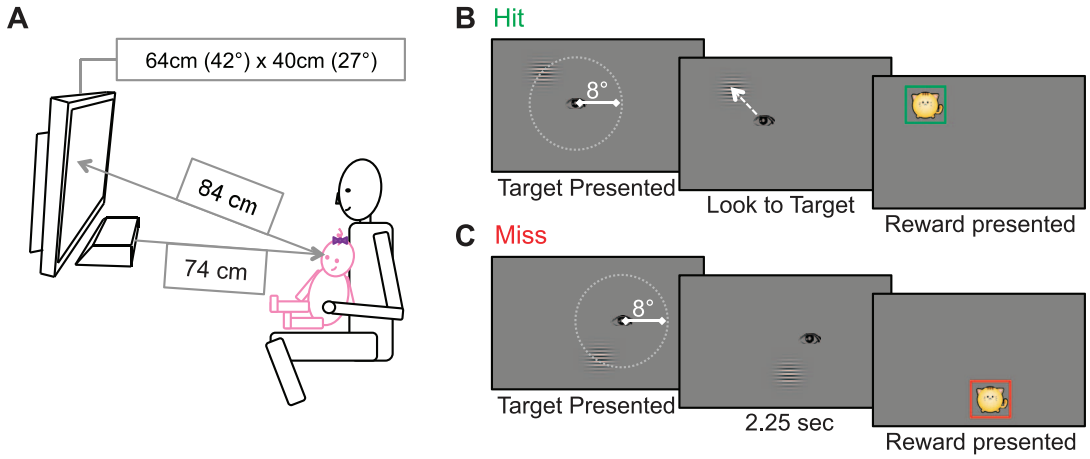

Figure 1. Empirical apparatus and procedure, reproduced with permission from Jones et al. (2014; ©ARVO). (A) The infant was seated on a parent's lap and viewed stimuli binocularly at a distance of 84 cm. A Tobii X120 eye tracker was mounted below the LCD screen and recorded the infant's eye movements. (B) In each trial, the screen was initially blank. A single Gabor grating was presented, which occupied approximately 3% of the total screen area. The center of the Gabor was located at a random location, 8° of visual angle from the infant's initial point of fixation. Infants had 2.25 s to look to the target (hit); otherwise, the trial was scored as a miss. (C) A visual "reward" was presented at the stimulus location at the end of every trial (independent of performance).

for optimal test design. Based on these analyses, we suggest simple maxims for improving test efficiency and demonstrate that many of the clinical practices outlined above (e.g., small numbers of trials, targeting a low threshold level of percentage correct performance, simplified computation of threshold) are logically sound. The findings may be valuable for increasing test efficiency when rapid threshold measurements are required in both clinical and basic research settings.

Methods

Empirical data

Empirical data were obtained using an automated test of infant visual acuity, the final version of which was reported in Jones et al. (2014). Note that the present work is primarily concerned with measurements of infant detection thresholds in general. The specific task used to provide empirical data and the variable that it measured (visual acuity) are relatively unimportant, and we believe that qualitatively similar results could be derived using other infant sensory detection tasks.

Participants contributing to the present data set were 55 infants, aged 2.6–12.7 months. Partial data for 30 infants were previously reported in the cited publication along with a detailed account of the methods. The research was carried out in accordance with the Declaration of Helsinki and was approved by the local NHS England ethics committee.

The "task" was to detect a high-contrast Gabor grating (Figure 1). Infants viewed a monitor on which a high-contrast black-and-white grating was presented against an isoluminant background. Gratings with a spatial frequency too high to be resolved would be invisible. Conversely, gratings with a spatial frequency below threshold would be visible, and infants were expected to fixate visible gratings, given their preference for pattern over uniformity (Fantz, 1958). The spatial frequency of the grating was manipulated between trials in order to measure visual resolution acuity. In each trial, the grating appeared at a random location on the screen, centered anywhere along the circumference of an 8° diameter circle (which was, in turn, centered on the infant's point of fixation at trial onset). Infants responded by fixating the grating (±3.6° of its center). Eye movements were recorded using a Tobii X120 remote eye tracker (Tobii Technology, AB) and were scored automatically as "hits" or "misses" by a computer-based algorithm (see Jones et al., 2014, for details).

Responses were classified in terms of two alternatives: Did the infant look at the target (hit) or not (miss)? In terms of scoring, the task could be described therefore as having "two alternatives." Alternatively, in adult psychophysics, tasks are often described in terms of the number of response alternatives available to the participant (for an overview of task nomenclature, see Kingdom & Prins, 2009). However, such an approach is inappropriate for the present task because the infant was not constrained to behave in 1 of M ways. Thus, an infant could miss (fail to fixate) the target either by making no eye movements or by looking at any number of nontarget locations (on or around the screen). An

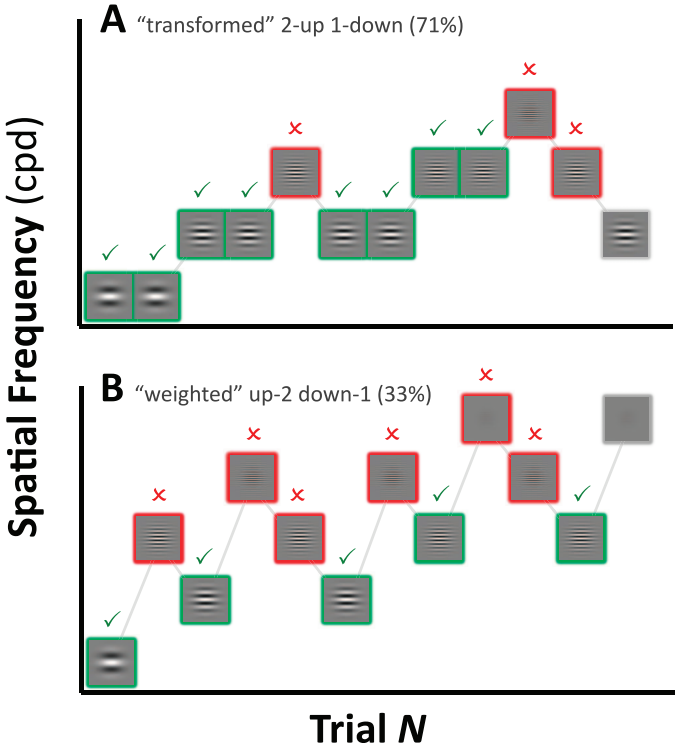

Figure 2. Example (A) transformed two-up, one-down staircase, and (B) weighted up-two, down-one staircases. (Note that in the present paper, "up" corresponds to "harder.") Correct and incorrect responses are highlighted by green ticks and red crosses, respectively. In the transformed staircase, the unit step size was of unit length. Two correct answers were required to increase difficulty whereas only one incorrect answer was required to decrease difficulty. This staircase will converge on a threshold of 70.7% correct. In the weighted staircase, upward steps were twice as large as downward steps, and a step occurred after every answer (correct or incorrect). This staircase will converge on a threshold of 33.3% correct.

important corollary of this is that, because the infant cannot be required to select from a discrete set of responses, it is impossible to know a priori the likelihood of a correct response occurring by chance (i.e., when the infant does not detect the target). For example, an infant who searches actively for an unseen target would tend to be more likely to look at the target location than an infant who only makes eye movements in response to seen targets. Guessing rates on non-forced-choice tasks must therefore be quantified empirically as detailed below.

Although the task was identical for all infants, the parameters of the adaptive staircase were varied between two cohorts. Twenty-five infants performed a two-up, one-down *transformed* staircase (Levitt, 1971), targeting 71% correct (see Figure 2A). The remaining 30 infants performed an up-two, down-one *weighted* staircase (Kaernbach, 2001), targeting 33% correct (see Figure 2B). Note that we here use "*N*-up/down" to refer to paradigms in which *N* trials are required to produce one adaptive step (transformed staircase) and "up/down-*N*" to refer to situations in which *N* steps are made after each single trial (weighted staircase). Also note that in the present paper, "up" always corresponds to "harder" (higher spatial frequency) although in many other tasks this relationship is reversed (e.g., when discussing luminance detection or sound pressure level detection).

Each infant performed the same staircase between one and three times ($\mu = 2.4$), depending on for how long the infant remained attentive. The staircase began at 0.88 cpd. The unit step size was 0.5 octaves. The floor was 0.61 cpd. The ceiling was 15.3 cpd, after which point no target was displayed (blank trial). The adaptive staircase continued until at least 15 trials and two reversals had occurred. Three catch trials followed the main procedure: One of these was blank (invisible target), and two were suprathreshold (0.61 cpd, a stimulus level that was easily visible for all infants tested—visibility was confirmed by previous normative data and by the empirical estimates of threshold reported here). These catch trials did not affect threshold estimates but allowed guessing rates (probability of answering correctly by chance alone without having seen the stimulus) and lapse rates (probability of not responding to a normally visible stimulus) to be quantified. Thus, failure to fixate a suprathreshold target catch trial was taken to indicate a "lapse," and fixation of the target location on an invisible target catch trial was taken to indicate a lucky guess. Note that, in keeping with the psychophysical literature (e.g., Prins, 2012; Wichmann & Hill, 2001), we use the term "lucky guess" loosely to describe any situation in which a response was not determined by the stimulus but was nonetheless scored as correct (e.g., including when the stimulus was not visible or when a lapse occurred). It should not, however, be taken to imply that infants were actively attempting to find unseen targets. Thus, a correct "guess" could occur by chance if, for example, the infant happened to look in the direction of the stimulus as he or she turned to face a parent. Note also that lapse and guess rates are determined both by the characteristics of the infant and any misclassification of responses by the experimenter/eye tracker. However, in the reported data, eye movements were classified using an automated eye-tracking algorithm, and anecdotally, classification errors occurred only rarely (see Jones et al., 2014).

It is important to stress that talk of lapses or guesses should not be taken to imply any particular behaviors on the part of the infant. For example, infants may make a false negative (lapse) response either by making no response (e.g., because they are tired) or by making an incorrect eye movement response (e.g., because they are distracted by something extraneous). The paradigm

did not distinguish between these different types of behaviors and simply scored infants depending on whether the infant fixated the target location.

Simulated data

Simulations were also used to explore how various adaptive tracking algorithms would be expected to perform given the challenges particular to testing infants (high lapse rates, few trials). Except when stated otherwise, simulated guess rates and lapse rates were set to their group-mean empirical values as observed in the 30 infants who performed catch trials. Other parameters, such as the shape and slope of the psychometric function, were fixed at values representative of the infant visual acuity literature (Teller, Mar, & Preston, 1992). Note that these latter values were only intended as plausible approximations, and small deviations would not have affected the present results qualitatively.

Simulated observers followed the principles of signal detection theory (SDT; Green & Swets, 1974; Macmillan & Creelman, 2005; Wickens, 2002). In each trial, there was a probability, λ, that the simulation had an "attentional lapse" (the lapse rate; Green, 1995; Prins, 2012; Wichmann & Hill, 2001). When such lapses occurred, the simulation made an unbiased guess, which had a fixed probability of being correct, γ, independent of the stimulus level (the guess rate). Otherwise, the simulation responded by comparing the internal estimate of the stimulus, x, to a fixed decision criterion (which was assumed to be ideal/unbiased throughout). Following SDT, the internal estimate was a noisy but unbiased representation of the true stimulus level. Specifically, in each trial, x was the sum of the true stimulus value and a value drawn from a zero-mean, log-normal random distribution intended to represent the "internal noise" inherent in observers' decision making. In practice, this model of decision making meant that the probability of responding correctly was determined by a log-normal psychometric function, thus

$$P(correct|x) = \begin{cases} \gamma, & \text{if } R \sim U([0,1]) < \lambda \\ (\gamma-1)(\lambda-1) & \\ \quad \times (1 - \ln\Phi(x;\mu,\sigma)) + \gamma, & \text{otherwise} \end{cases}, \quad (1)$$

where $R \sim U([0, 1])$ is a random value uniformly distributed between 0 and 1, and $\ln\Phi$ is the log-normal cumulative density function:

$$\ln\Phi(x|\mu,\sigma) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left[\frac{\ln x - \mu}{\sqrt{2}\sigma}\right]. \quad (2)$$

Example functions corresponding to Equation 1 are shown graphically in the Results (Figure 3B). Note that the use of logarithmic noise merely reflects how psychometric functions have been observed to be distributed on grating-detection tasks (e.g., Anderson, Evans, & Thibos, 1996). When plotted on a log-transformed x-axis, Equation 1 produces a cumulative Gaussian sigmoid.

Results

Lapse rates are higher in infants, but rates of guessing correctly are generally lower

To characterize infants' lapse rates, λ, and guess rates, γ, the three catch trials at the end of each test sequence were examined. Aggregating across infants, lapse rates (failure to fixate a suprathreshold target) were 25.9%, and guess rates (fixating the target region in a blank trial by chance) were 6.9%. From this, one would expect infants to respond independent of the stimulus on one in four trials and to be scored correct by chance in one in 14 of these trials.

From inspection of Figure 3A, it can be seen that lapse and guess rates were stable across repeated test runs. Furthermore, error rates were broadly similar across younger (3–6 months) and older (6–12 months) infants, although we cannot speculate on infants outside this range, and there is likely to be substantial individual variation between infants irrespective of age.

The error rates observed in infants are markedly different from those found in typical, "laboratory" psychophysics with either infants or adults. For example, in a classic 2AFC task, the expected guess rate is 50%, and lapse rates tend to be almost zero. This corresponds to a psychometric function in which proportion correct varies sigmoidally between approximately 0.5 and 1.0 (Figure 3B, dashed line). Such functions are commonly reported for studies using 2AFC methods in both adults and infants (Gwiazda, Wolfe, Brill, Mohindra, & Held, 1980; Teller et al., 1982). In contrast, infants on the present clinical task were far more likely to miss a visible target (lapse) and far less likely to be correct by chance (guess).

The infants' low guess rate, γ, reflects a low likelihood that a response would be scored as a "hit" by chance. In the present test, low guess rates were, in part, due to properties of the methodology (a large number of possible target locations and a score of "miss" if the infant failed to fixate the target within a fixed time frame) and were, in part, due to properties of the infant participants (an indisposition to search actively for an unseen target). Notably, the large number of target locations was made possible by the

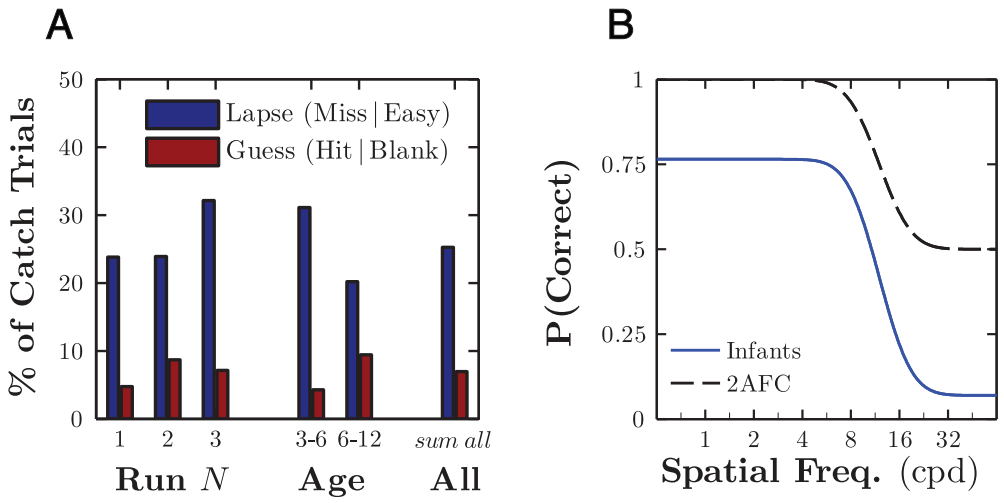

Figure 3. (A) Group aggregate false negative (lapse) and false positive (guess) rates for suprathreshold target catch trials and invisible target catch trials, respectively. These error rates are shown: (a) broken down by experimental block number, (b) broken down by infants' age (in months), and (c) aggregated across all blocks/infants. Note that the guess rate describes the likelihood of a "guess" being correct, not of a guess occurring. (B) Example model psychometric functions for an "infant" observer with a guess rate and lapse rate corresponding to the (sum all) empirical infant data (blue) and for an idealized "adult"-like observer performing a 2AFC task with zero lapse rate (black, dashed). The slope and mean of these functions were chosen arbitrarily for display purposes.

use of an eye tracker to register precisely where the infant was looking. In contrast, human coders are considerably less precise at judging gaze location,[2] and beyond two target locations, any improvements in measurement sensitivity tend to be outweighed by increased trial durations and greater classification errors (Teller, 1979). Thus, when using a human coder to score infants' responses, it might be expected that the guess (false positive) rate would be much greater than reported here. However, this need not necessarily be the case. See General discussion for more regarding strategies for reducing guess rates.

Infants also exhibited high lapse rates, λ. This reflects their relative inattentiveness and the lack of an explicit task. Adult observers in traditional psychophysical paradigms tend to be diligent, motivated, and task-oriented. As a result, responses are determined almost exclusively by task-relevant perceptual information. In contrast, infants are under no imperative to perform the experimenter's task, and responses may be influenced by many variables, including level of motivation, interest, extraneous events, and outcomes of previous trials as well as by task-relevant perceptual factors. It is therefore unsurprising that infants often failed to fixate a supra-threshold stimulus even when it was clearly visible to them. Interestingly though, it is also common, even for laboratory studies of *infants*, to report psychometric functions that asymptote near one, implying near-zero lapse rates (e.g., Gwiazda et al., 1980; Teller et al., 1982). This may reflect a systematic difference between clinical and laboratory testing with the latter incorporating less distracting environments or more compliant participants. However, a more parsimonious explanation is that differences in lapse rates are an artifact of how data are collected and reported. Thus, in the present experiment, stimuli were presented based on a predetermined, automated algorithm, and every stimulus presentation was considered a (valid) trial when analyzing the data. However, as discussed in the General discussion, in many laboratory experiments, testing will be paused if an infant appears inattentive, and trials will be excluded, post hoc, if the infant did not appear to be paying attention. In this way, the *effective* lapse rate present in the reported data may be near zero even if the infant was not paying attention for a substantial proportion of the time. Notably though, such an approach is inappropriate for rapid assessments (i.e., when the opportunity to repeat trials is limited) or when using automated methods in which lapses in attention cannot be identified reliably.

Given low guess rates and high lapse rates, the ideal threshold to target is often $\leq$50% correct

When using adaptive staircases, most authors recommend targeting a high performance threshold, such as 83% correct (McKee et al., 1985) or 94% correct (Green, 1990; Klein, 2001; although cf. García-Pérez, 2001). However, as McKee et al. (1985) acknowledge, high target thresholds cease to be appropriate once lapse rates increase beyond a few percentage points. In fact, given the 26% lapse rate observed in the first section of Results (Figure 3), thresholds above 74% correct would not even exist on the infant psychometric

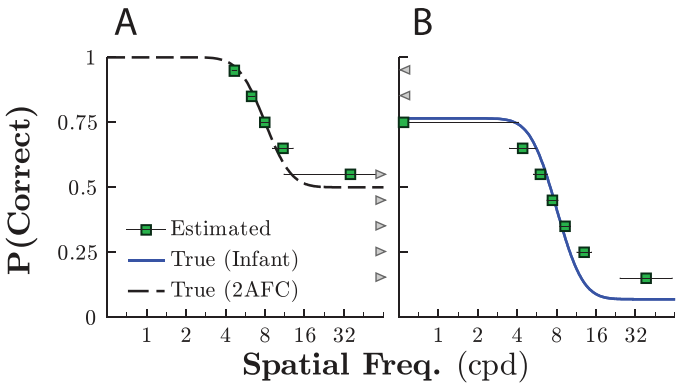

Figure 4. Effect of target performance level on threshold estimate accuracy. Mean (±1 *SD*) threshold estimates (square symbols) were made at 10 levels of target percentage correct. Mean threshold was computed by averaging more than 2,000 independent simulations. Within each simulation, a threshold was computed by averaging over the last 64 reversals of a 256-reversal staircase. The continuous distributions show the true psychometric function for (A) an idealized 2AFC observer performing a 2AFC task (with zero lapse rate), and (B) a simulated infant participant with a lapse rate and guess rate corresponding to the empirical values observed in the first section of Results (see Figure 3). Note that error bars for thresholds estimated near the upper and lower limits of the function extend beyond the plotted range. Gray arrows denote threshold estimates that fell outside of the displayed range (e.g., staircases targeting thresholds lower than the guess rate tended upward toward infinity).

function. Targeting them would result in threshold estimates that tended upward toward infinity (or until the measurement ceiling was reached).

Lower targets, such as the commonly used threshold of 71% correct, does exist on the infants' psychometric function but would not be ideal given the lapse rates and guess rates observed in the first section of Results. Thus, Figure 4 shows the results of Monte Carlo simulations in which differently weighted staircases were used to target various points on the psychometric function (Kaernbach, 1991). The results are clear and unsurprising. When guess rates were nonzero, as in the case of an idealized 2AFC observer (Figure 4A), thresholds near floor tended to be overestimated. Conversely, when lapse rates were nonzero, as in our infant data set (Figure 4B), thresholds near ceiling tended to be underestimated. Furthermore, for these "borderline" cases, the level of measurement variability was high; repeated thresholds often fluctuated by an octave or more even though each estimate was computed using many hundreds of trials (see Figure 4). The ideal target for providing an unbiased estimate of sensory ability lay at approximately the midpoint of the function. This midpoint would be at 75% correct for an idealized adult performing a 2AFC task but was 40.5% given the empirical infant data in the first section of Results. Thresholds in this region can be tracked, for example, by using a one-up, one-down staircase to target 50% correct or by using a weighted up-two, down-one staircase to target 33.3% correct.[3]

In short, Figure 4 shows how thresholds estimated far from the midpoint of the psychometric function become progressively inaccurate and imprecise. Moreover, extreme threshold targets also have additional undesirable consequences for researchers looking to perform statistical comparisons between groups of infants. For example, in the Supplemental Material, we show how, when targeting an extreme value of percentage correct performance, interindividual variability in lapse or guess rates can lead to a highly skewed sampling distribution of threshold estimates. This can cause summary statistics, such as group-mean threshold, to become misleading and complicates the use of parametric statistical tests such as *t* tests and ANOVAs.

Speed is also critical when considering what performance level to target because infants often provide only a small number of trials (and as will be noted in the third section of Results below, the *effective* number of trials may be lower still). It is therefore imperative that the staircase converges on threshold as quickly as possible.

Figure 5 shows simulations demonstrating how number of trials until convergence varies as a function of the staircase algorithm and as the distance from true threshold in trial one varies (Figure 5A). When the staircase begins near threshold, speed is of trivial concern; virtually any staircase algorithm will yield estimates close to the observer's true threshold in around a dozen trials (Figure 5B). However, when true threshold is far from the starting point, the situation is more complex. For example, in the present task, the acuity threshold of a 1-year-old infant was expected to lay approximately nine steps from the initial stimulus value (given a starting level of 0.88 cpd). In this case, a staircase targeting a higher level of percentage correct would be expected to take substantively *longer* to converge (Figure 5C; e.g., solid black line vs. gray dotted line). This is counterintuitive as stimuli corresponding to higher percentage correct performance lie *closer* to the starting point. However, the difference in convergence rate is explained by the present combination of high lapse rates and low guess rates. Together, these result in frequent chance misses (false negatives) but relatively few chance hits (false positives). In staircases that target thresholds above 50% correct, steps following misses are larger and/or occur more often than steps following hits. Such staircases are therefore susceptible to becoming skewed downward (underestimation of threshold). By contrast, in a weighted up-two, down-one (33.3%) staircase (see

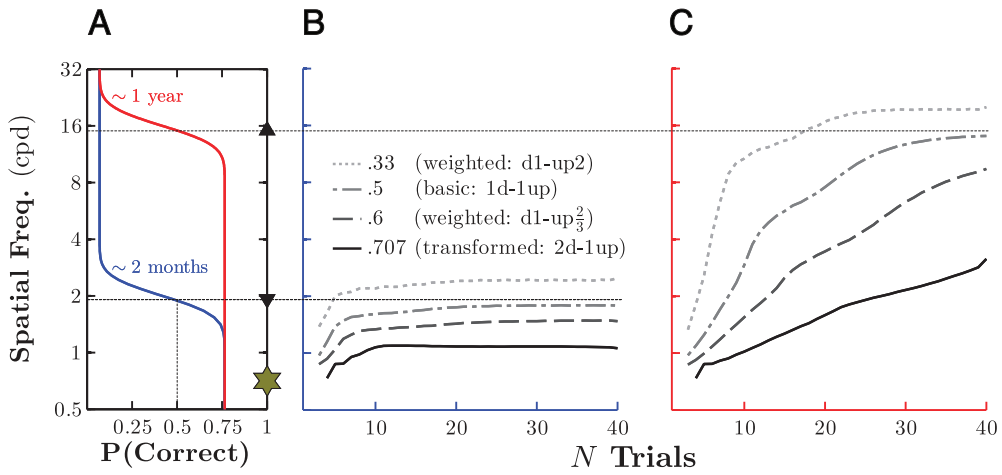

Figure 5. Estimated threshold as a function of number of trials, using four different threshold targets and either a low or a high underlying threshold. (A) The underlying psychometric functions for simulated young and old infants. Markers on the right show the starting point of all adaptive staircases (gold star) and the true 50% correct threshold for the younger (down arrow) and older (up arrow) infant. For clarity, the true values for the other three target thresholds are not marked but can be inferred by reading upward from the x-axis and then across. (B) Results of 20,000 Monte Carlo simulations (per staircase algorithm) for the simulated young infant, showing how threshold estimates made using each of four staircase algorithms vary with number of trials. In each case, threshold was computed as the mean of the highest (even) number of reversals available (C) Same as (B), for the simulated older infant (i.e., more steps required to reach true threshold given same staircase starting point).

Figure 2B), steps following chance misses have less impact on the staircase because a false negative response can be compensated for by a hit on either of the two subsequent trials. Staircases targeting lower thresholds are therefore more robust to false negative results (although, on the other hand, they would be more susceptible to becoming skewed upward by false positive results, i.e., if both lapse rates and guess rates were high).

Considerations of speed therefore lead to the same conclusion as those of accuracy: Given high lapse rates and low guess rates, targeting a threshold of around 40.5% is preferable to targeting the higher levels of performance recommended in adult psychophysics.

We assessed these findings empirically by comparing infant acuity thresholds (measured in two independent cohorts of infants as described above), using either a weighted up-two, down-one staircase (33.3%; $N = 30$), or a two-down, one-up transformed staircase (70.7%; $N = 25$). The results are shown in Figure 6. The staircase targeting 33.3% correct (Figure 6, circles) provided robust visual acuity estimates. There was a strong improvement in acuity with age, $F(1, 28) = 35.06$, $p < 0.001$, and this trajectory replicated, $\beta_{obs} = 0.98$ versus $\beta = 1.0$, $t(28) = -0.38$, $p = 0.710$, *n.s.*, previous reports that visual acuity improves by 1.0 cpd/month within the first year of life (Salomão & Ventura, 1995; see also Dobson & Teller, 1978; Mayer et al., 1995). In contrast, the 70.7% correct acuity estimates were markedly poorer. Thresholds were consistently near floor and failed to yield any effect of age, $F(1, 23) = 1.95$, $p =$

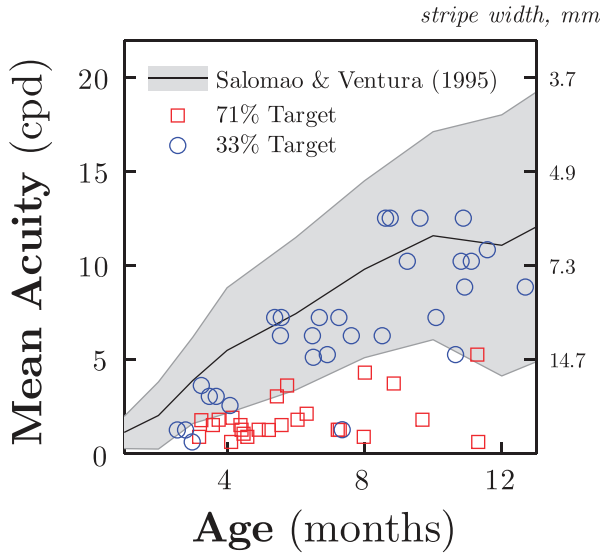

Figure 6. Empirical acuity thresholds, measured using either a high (red squares) or low (blue circles) threshold target. Each point represents the geometric mean of one to three threshold estimates, averaged within a single infant. The gray shaded region shows the 90% tolerance limits for normative data ($N = 646$) from Salomão and Ventura (1995). The blue circles show good-accuracy data, derived using a weighted up-two, down-one staircase and reported previously in Jones et al. (2014). The red squares show low-accuracy data, derived using a transformed two-down, one-up staircase (previously unreported).

0.176, *n.s.* The empirical data therefore support the conclusion that targeting a low percentage correct threshold is liable to yield a more accurate estimate of an infant's perceptual sensitivity.

Given low guess rates and high lapse rates, max correct can provide the most robust measure of threshold

The textbook way to compute threshold from an adaptive staircase is to average the variable parameter over the last N reversals (i.e., the last N occasions on which the stimulus changed from increasing to decreasing in magnitude or vice versa). In practice though, many clinical tests simply take the maximum correct response (i.e., the lowest stimulus magnitude that the subject was judged to see or hear) as the measure of an infant's threshold (e.g., Day et al., 2008; Teller et al., 1986). To investigate the validity of the max correct rule, Monte Carlo simulations were run, comparing the effects of basing thresholds simply on max correct versus averaging over two, four, or eight reversals. The simulated observer's properties corresponded to those shown graphically in Figure 4B. The staircase always targeted 33.3% correct, and its parameters were identical to those described in the empirical methods (see also Figure 2B).

The results of these simulations are shown in Figure 7. In terms of the number of trials required to estimate threshold accurately (Figure 7A), the eight-reversal method was noticeably slower. It failed to converge reliably until at least 20–25 trials, which is too long for some pediatric tests. However, there was little difference between using a smaller numbers of reversals and/or with using the max correct rule; all of these methods produced similar mean threshold estimates when trials were few. As the number of trials increased beyond 25, threshold inflation started to become problematic for the max correct rule (i.e., because a single lucky guess can permanently raise the estimate). However, this is likely outside the range of many clinical tests. For example, the infants in the second section of Results only completed 15 trials per staircase, on average.

Second, we considered the number of trials required to yield *any* estimate of threshold (Figure 7B). Here again, there was a clear difference between using eight reversals and the other techniques. It took 23 trials for eight reversals to be guaranteed whereas using fewer reversals or a max correct rule could guarantee an estimate of threshold by around five to 10 trials.

Third and finally, we examined variability between estimates (Figure 7C). Unsurprisingly, averaging over more reversals gave a more reliable estimate of threshold (although this may not hold in practice; see below). But again, there was relatively little difference

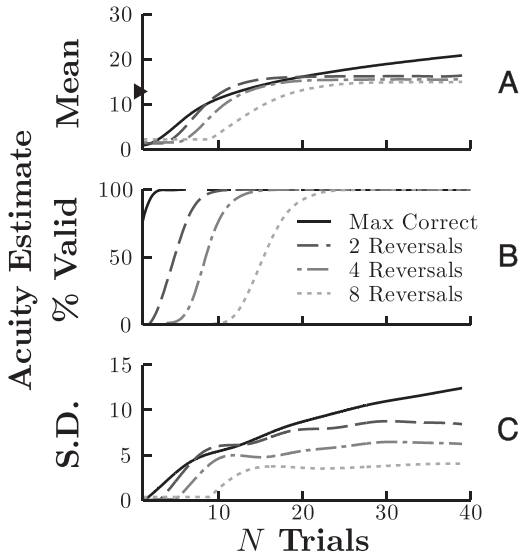

Figure 7. Effects of threshold computation method on estimated acuity as a function of number of trials. Data were derived using Monte Carlo simulations ($N = 20{,}000$) (A) Mean threshold estimate (true value given by the triangular marker on the ordinate). (B) Proportion of staircases yielding a threshold estimate. (C) Standard deviation of threshold estimates, smoothed using polynomial spline fitting.

in standard deviation between max correct, two, and four reversals.

In short, the simulations appeared to show that, with few trials ($N < 25$), there is relatively little difference between using max correct and two or four reversals and that averaging over more reversals gives more reliable estimates but at the risk of not yielding any estimate at all within the number of trials permitted.

These conclusions were largely supported by the empirical data. Thus, Table 1 shows the results when the four methods of threshold computation were applied to the empirical data ($N = 74$ staircases, from the 30 infants who performed the 33.3% target condition only). As the simulations predicted, threshold estimates were very similar across max correct and two or four reversals. Conversely, requiring eight reversals resulted in results that were correlated overall but which exhibited lower average thresholds and less estimate variability but a larger number of staircases being excluded (for failing to yield sufficient reversals to compute a threshold). Exclusion rates were approximately uniformly distributed across the three staircases ($S_1 = 19\%$, $S_2 = 24\%$, $S_3 = 14\%$) and between infants.[4]

Based on the forgoing, it appears that the best way to compute threshold might be to average over the highest number of reversals available in a given staircase.[5] However, on close inspection, it can be seen that averaging over small numbers of reversals actually resulted in slightly *greater* test–retest variability than

Method	*N* excluded	ΔT	*r*	*p*	|Test–retest|
Max correct	0 (0%)	n.a.	n.a.	n.a.	0.82
Two reversals	0 (0%)	0.04	0.90	<0.001	0.85
Four reversals	1 (1%)	−0.03	0.92	<0.001	0.92
Eight reversals	42 (57%)	−0.09	0.88	<0.001	0.57

Table 1. Effects of threshold computation method on empirical threshold estimates. *Notes*: From left to right: method of threshold computation; number of staircases excluded (data derived from 30 infants; $\mu = 2.4$ staircases per infant); mean difference in estimated threshold in octaves (negative = max correct was higher); *r* and *p* values for correlations with max correct; mean absolute within-infant test–retest difference in octaves (higher = greater threshold estimate variability). Test–retest differences were computed using successive staircase: S_1 versus S_2, S_2 versus S_3 (each infant was tested two to three times within a single ~30-min period with short breaks between successive staircases).

the max correct method (Table 1, column 6). This was unexpected given the simulation results, which had suggested that statistical reliability improves monotonically with number of trials (Figure 7C). Furthermore, inspection of the individual data indicated that on some occasions, averaging two or four reversals produced thresholds that were *markedly* lower than either the max correct estimate or what would be expected given the infant's age (Figure 8, crosses).

To see why greater numbers of reversals may not provide more reliable estimates of threshold, consider that to average over multiple reversals is to tacitly assume that "the psychometric function is stationary over time" (Levitt, 1971, p. 468), that is, to assume that each stimulus level will always produce the same expected level of performance. Although stationarity may be an acceptable assumption in adults (although fatigue, learning, or bias make it never strictly true; see Jones, Shub, Moore, & Amitay, 2015), it is often patently false in infants, whose behavior often varies markedly even within a short testing session. For example, some infants may be inattentive at the beginning of a test (e.g., reaching for the rattle that they were previously playing with) but then settle. Other infants may perform well initially but then become agitated or bored. These changes may happen suddenly or gradually over time. Moreover, the loss of attention may be permanent, resulting in a "V"- or "Λ"-shaped staircase, or it may fluctuate (for example in response to hitting threshold), resulting in a "sawtooth" staircase. Both of these patterns were observed in individual infants among those tested in the second section of Results (Figure 9A). Furthermore, as has been previously reported (Atkinson, Wattam-Bell, Pimm-Smith, Evans, & Braddick, 1986), changes in attention are particularly great in infants older than 6 months, who are often very alert initially but are liable to quickly lose interest.

The corollary of this is that it is not always true that "the longer the staircase, the more accurate the estimate" (García-Pérez, 2001). In some cases, estimates may not improve. For example, Atkinson et al. (1986) found no significant difference between acuity estimates made based on 20 trials and those using 50

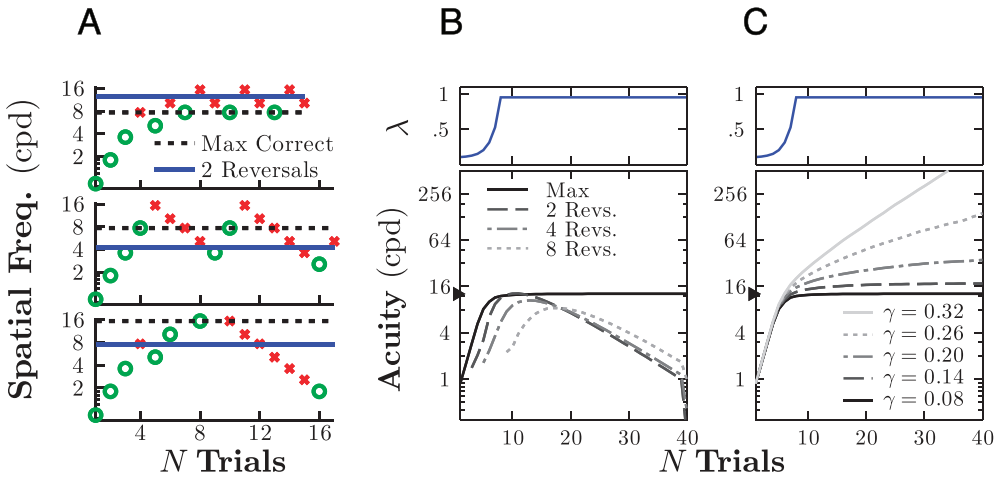

Figure 8. Individual differences between thresholds computed using a max correct (x-axis) versus a two-reversals (y-axis) rule. Each point represents a threshold computed both ways using a single infant's data (same empirical data used to compute both thresholds). Each infant contributed either two or three data points (i.e., each infant performed two to three adaptive staircases). Points at which the max correct and two-reversal estimates deviate by more than one octave are highlighted with a red cross.

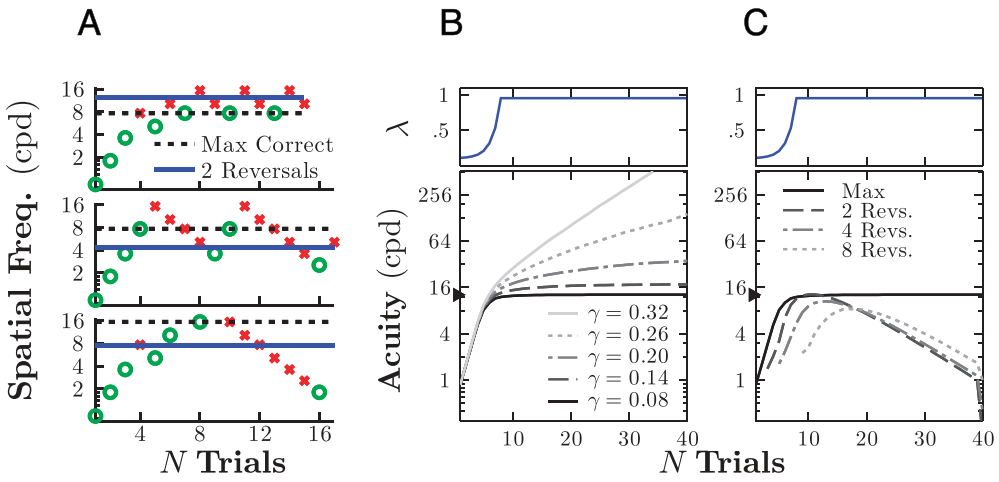

Figure 9. Effect of nonstationary inattentiveness. (A) Example adaptive staircases from three infants, demonstrating a "compliant" staircase (top: observer attentive throughout), a "sawtooth" staircase (middle: transient loss of attention), and a "Λ"-shaped staircase (bottom: permanent loss of attention). Horizontal lines indicate the resultant threshold estimates using either max correct (dashed) or two reversals (solid). (B) Monte Carlo simulations ($N = 20{,}000$), demonstrating how an increase in lapse rates (inattentiveness) over time can affect threshold estimates given various thresholding strategies. (C) Monte Carlo simulations, demonstrating how the max correct rule is liable to produce inflated threshold estimates as the guess rate, γ, increases.

trials. In other cases, longer staircases may produce *less* accurate estimates. For example, if an infant gradually loses interest in the task, then averaging over the final reversals will cause the infant's ability to be systematically underestimated (Figure 9B). In contrast, the max correct rule does not require any assumptions about *when* the infant was or was not paying attention, only that he or she performed the task consistently at some point. In this sense, max correct may be a more robust measure of threshold when fluctuations in attentiveness are a concern and/or when no independent means exist by which to identify when such lapses have occurred.

It is important to stress, however, that the benefits of the max correct rule only hold when the task is designed to achieve low guessing rates. In contrast, when the probability of a false positive response is high, then an inattentive participant is liable, purely by chance, to produce a run of correct responses. Under the max correct rule, this would cause estimates of perceptual sensitivity to be overestimated permanently thereafter. As shown in Figure 8C, a guessing rate of around 14% or less would be required to avoid spurious overestimates in the present task, given the expect number of test trials (~15–20).

General discussion

In adults, the question of how to measure psychophysical thresholds quickly, accurately, and reliably has been the subject of numerous studies (e.g., Amitay, Irwin, Hawkey, Cowan, & Moore, 2006; L. G. Brown, 1996; García-Pérez, 2001; Green, 1990; Gu & Green, 1994; Johnson, Chauhan, & Shapiro, 1992; Klein, 2001; McKee et al., 1985; Rose, Teller, & Rendleman, 1970; Saberi & Green, 1996; Watson & Pelli, 1983). However, relatively few authors have investigated the challenges particular to working with infants, such as the (a) very limited numbers of trials, (b) high variability between individuals, and (c) high levels of participant inattentiveness (cf. Banks & Dannemiller, 1987; A. M. Brown, 1990; García-Pérez, 1998; Swanson & Birch, 1992; Teller, 1979; Viemeister & Schlauch, 1992; Werner & Marean, 1991). Furthermore, even fewer authors (Atkinson et al., 1986; Carney, 1992) have considered how to maximize test efficiency in a clinical context, in which reliability and speed are particularly important.

Here we used both simulations and empirical data to explore how to maximize test efficiency in infants. The first section of Results showed that a successful infant psychophysical test can be designed to offset low guess rates (false positives) against high lapse rates (false negatives). Given such task statistics, the second section of Results demonstrated that targeting a relatively low threshold (e.g., 33.3% correct) provides a faster and more accurate measure of ability than traditional, "adult" targets of 70%–94% correct. Finally, the third section of Results showed that when the guess rate is low, simply taking the maximum correct response is a viable way to compute threshold and is more robust to fluctuations in attentiveness.

Other approaches to improving to improving test efficiency

In the present work, we have focused on how to accommodate the statistical differences that exist between infant and adult psychophysics, in particular, the high lapse rates exhibited by infants. An alternative approach is to make infants respond in a more adult-like manner, thereby obviating the need for a special, "infant-friendly" psychophysical approach. For example, some experimenters have attempted to minimize inattentiveness by devising highly salient stimuli, such as visual acuity gratings arranged into the pattern of a smiling face (Harris, Hansen, & Fulton, 1984). More salient stimuli may be an effective means of engaging some infants and reducing lapse rates. However, their exact effectiveness is unclear, and they can seldom guarantee sustained attention in all infants. Furthermore, the ability to construct engaging stimuli is often limited by theoretical considerations, such as the need to minimize inadvertent response cues (e.g., edge effects; Campbell, Carpenter, & Levinson, 1969; Kelly, 1970). Alternatively then, some authors recommend the use of suprathreshold "reminder" stimuli. These may be introduced periodically (Atkinson et al., 2006) or after incorrect responses (e.g., Mayer et al., 1995) and are again designed to mitigate or prevent a loss of concentration, especially as the staircase approaches threshold (see Figure 9A). The precise effectiveness of this technique and its relative tradeoff with overall test duration remain largely unknown, however, and in our own work, such trials are generally used to identify inattentiveness rather than militate against it.

A related approach is to use an algorithm to judge when a child is being inattentive and to disregard those trials. Thus, a human experimenter will often exclude a trial post hoc if he or she feels the infant was not paying attention and will not start the next trial until the infant appears receptive. In this way, the *effective* lapse rate within "valid" test trials is reduced, sometimes to adult-like levels (e.g., Gwiazda et al., 1980; Teller et al., 1982). The drawback of this approach is that it is highly labor intensive and can easily lead to subjective biases if the adult experimenter is not blind to expected performance. Nonetheless, with an expert operator, the preemption and exclusion of inattentive trials can be a highly effective means of improving test efficiency. It remains an interesting and open question whether such approach can be effectively formalized within an automated system, however. In lieu of such a mechanism, the present suggestion of combining low guess rates with a low target threshold may therefore be of particular use to authors looking to standardize or automate the process of testing.

Finally, another potential way to improve test efficiency is to adopt a Bayesian approach in which prior knowledge is used to guide future threshold estimates. For example, as discussed in the second section of Results, most adaptive techniques work acceptably well when the staircase is made to start near true threshold. Accordingly, some infant testing procedures recommend varying the start point, depending on the age of the infant (e.g., Teller et al., 1986). However, this technique is liable to yield misleading results if the prior data are either inaccurate or inappropriate. If, for example, the question is whether the child's threshold is normal for his or her age, then a strong prior will bias the result toward the affirmative, increasing the chance of a type II error (misdiagnosis as normal). Moreover, appropriate normative data is not always available, especially when working with infants or patients.

The importance of low guess rates

The present results demonstrate that a key requirement for a successful test of infant detection threshold is a low guess rate. Thus, as stressed in the Results, the recommended use of low threshold targets (e.g., 33% correct) and a max correct rule only makes sense when the likelihood of making a correct response by chance is minimized.

In the empirical work reported here, low guess rates were obtained by using eye tracking to distinguish between a large number of possible target locations. However, high-precision technology is only one way to minimize false positive response, and human-coded tests can also exhibit surprisingly low guess rates. For example, anecdotal observations suggest that when a human operator "manually" classifies an infant's responses to a clinical Acuity Card stimulus, guess rates are also substantially below 50% despite the target always appearing in one of two locations. In the case of Acuity Cards, low guess rates are due to the fact that the coder is not actually making a forced-choice left–right gaze judgment but is instead making "a broadly integrative, subjective judgment" (McDonald et al., 1985, p. 158) of whether the infant saw the stimulus based on a wide range of information (e.g., the infant's eye movements, facial expression, pupil response, and any physical movements; Mayer et al., 1995). Another technique for reducing guess rates is to jitter the onset of the target stimulus randomly in time (much as the target was randomly jittered in space in the acuity test reported here). This further minimizes the chance that an infant will make a correct response by chance alone and is a strategy commonly used in both infant (Day et al., 2008) and adult audiometry (British Society of Audiology, 2011). In short, a range of approaches can be used to reduce guess rates, including the use of both advanced technology and/or the insights of an expert

human. Exactly how the guess rates achieved using more traditional "expert judgment" techniques compare with those reported here, achieved using eye tracking and a high number of target locations, remains uncertain. Furthermore, there are unresolved concerns that some techniques involving human judgments may make the dependent measure susceptible to operator bias (for discussion, see Teller et al., 1986). However, the phenomenon itself is clear: namely, that clinical tests of infants typically exhibit much lower guess rates than tests of either infants or adults using standard psychophysical techniques (e.g., 2AFC methods).

Given that many existing clinical tests exhibit low guess rates, it is perhaps unsurprising that some of them also follow the other practices recommended here. Thus, although their articulation is novel, many successful clinical tests—by accident or design—already exploit our recommendations of a low percentage correct threshold algorithm, combined with a max correct analysis rule. For example, Teller Acuity Cards (McDonald et al., 1985; Teller et al., 1986) have proven highly robust as a measure of visual acuity. In clinics, this test targets a relatively low threshold of 50% correct and uses a max correct rule to compute performance. Similarly, in infant audiometry, a weighted staircase is used to target 33.3% correct, and threshold is computed using a modified ("two out of three") max correct rule (Day et al., 2008; Widen et al., 2000). The present findings make explicit what makes these tests effective and could assist in the design of similarly effective tests in the future.

Generalizability of the present findings

In the present paper, we have focused on identifying procedures that increase test efficiency in infants. However, the same considerations may also apply more generally to instances in which there is a premium on estimating thresholds rapidly, including in normal adults. For example, the variations in attentiveness reported in the third section of Results are not unique to infants, and qualitatively similar staircase dynamics have also been observed in both children (Moore, Ferguson, Halliday, & Riley, 2008) and adults unaccustomed to psychophysical testing (Jones et al., in press). Although lapse rates in such instances tend to be smaller and more transient than those reported here in infants, they might nonetheless cause sensitivity to be underestimated systematically. The same principles of using low guess rates, low thresholds, and a max correct rule may therefore also be effective when measuring perceptual sensitivity in children or when performance is changing rapidly due to learning, adaptation, or short-acting interventions, such as transcranial magnetic stimulation.

Finally, it is important to note that in this paper we have concentrated on how best to assess sensory *detection* abilities in infants. For example, in the empirical data reported, we used eye tracking to measure infants' ability to fixate a salient visual acuity stimulus presented against a uniform background. We believe that the methods used and the theoretical recommendations presented would readily generalize to other measurements of auditory or visual detection (e.g., luminance, color, motion). A second major class of task, however, concerns *discrimination* (e.g., can infants distinguish between two faces, Pascalis, de Haan, & Nelson, 2002, or two regions of a display containing different levels of statistical regularity, Wattam-Bell, 1992). For such tests, the analyses discussed here for visual detection are also appropriate although error rates may differ when the "foil" regions are not blank. Furthermore, when designing discrimination tasks, there may be additional complications to consider, which fall beyond the scope of the present work. For example, what constitutes a "response" when presented with two visible stimuli (e.g., first look, dwell time, number of fixations)? It may be that in discrimination tasks the limiting factor for a test's effectiveness may not be the performance level targeted or how threshold is computed, but how accurately responses can be classified on a trial-by-trial basis. In that respect, it remains unknown to what extent the process of response classification can be optimized and, in particular, the extent to which new technologies, such as remote eye tracking, can help to maximize speed, accuracy, and precision in tests of perceptual discrimination.

Conclusions

1. The task statistics of a recent test of infant acuity were shown to reflect a high rate of inattentiveness (26%) but a low rate of false positive guessing (7%). Moreover, inattentiveness was shown to be nonstationary with some observers losing interest throughout the course of testing.
2. Given these properties, it was shown that accurate, reliable, and fast estimates of threshold could be best achieved by using a weighted staircase to target a low threshold and by taking max correct performance as the index of perceptual sensitivity.
3. These recommendations were supported by empirical data. A transformed staircase targeting a high threshold (70.7% correct) underestimated performance whereas a weighted staircase targeting a low threshold (33% correct) yielded much more accurate results. Furthermore, a max correct rule was more

robust to fluctuations in attentiveness than averaging over reversals.

Keywords: psychophysics, infants, adaptive staircase, thresholds, signal detection theory

Acknowledgments

The authors thank John Wattam-Bell for helpful discussion and criticism. This work was supported by Fight for Sight, the National Institute for Health Research Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, the Special Trustees of Moorfields Eye Hospital, and the Leverhulme Trust.

Commercial relationships: none.
Corresponding author: Pete R. Jones.
Email: p.r.jones@ucl.ac.uk.
Address: Institute of Ophthalmology, University College London, London, UK.

Footnotes

[1] In the clinical Acuity Card procedure, a trained operator presents the infant with a sequence of cards, each containing a black-and-white grating on either the left or right side. Gratings vary in spatial frequency and are presented against a gray background of matched mean luminance. Given their preference for pattern over uniformity, infants will tend to fixate the grating pattern if they can resolve it. The operator judges whether the infants fixate the grating and determines the highest spatial frequency that they fixate reliably.

[2] The 95% confidence interval for observer estimates of gaze location is approximately ± 5.6° (Cline, 1967), versus the manufacturer's figure of ± 0.3° given the present, Tobii X120, eye tracker (Tobii Technology, 2012).

[3] More exactly, a threshold of 40.5% can be assessed by using a weighted up-1.5, down-one staircase to target a 40% correct threshold.

[4] When attempting to compute eight-reversal thresholds, four infants contributed the maximum of three excluded staircases, and five infants contributed the minimum of zero excluded tracks. The remaining 30 tracks were distributed among the other 21 infants. Exclusions were not due to just a small number of infants and were qualitatively consistent with a Poisson distribution.

[5] In the present data, staircases contained a variable number of reversals because infants were required to complete a minimum number of trials (and so could exhibit more than the minimum number of reversals). In other studies, a variable number of trials may also occur because a staircase was continued for as long as the infant was deemed to be attentive. Note that, in the present design, a trial could not continue if the infant was not attending to the stimulus display screen (i.e., if the eye tracker was not able to detect consistent fixations or detected substantial head movement). If the infant were to become uncooperative, then the prescribed number of trials could therefore not occur, and the staircase would be aborted. Such a situation never occurred within the data reported here, however.

References

Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., & Moore, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *The Journal of the Acoustical Society of America, 119*(3), 1616–1625.

Anderson, R. S., Evans, D. W., & Thibos, L. N. (1996). Effect of window size on detection acuity and resolution acuity for sinusoidal gratings in central and peripheral vision. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 13*(4), 697–706.

Armstrong, V., Maurer, D., Ellemberg, D., & Lewis, T. L. (2011). Sensitivity to first-and second-order drifting gratings in 3-month-old infants. *I-Perception, 2*(5), 440–457.

Atkinson, J., Braddick, O., Rose, F. E., Searcy, Y. M., Wattam-Bell, J., & Bellugi, U. (2006). Dorsal-stream motion processing deficits persist into adulthood in Williams syndrome. *Neuropsychologia, 44*(5), 828–833, doi:10.1016/j.neuropsychologia.2005.08.002.

Atkinson, J., Wattam-Bell, J., Pimm-Smith, E., Evans, C., & Braddick, O. J. (1986). Comparison of rapid procedures in forced choice preferential looking for estimating acuity in infants and young children. In *Detection and measurement of visual impairment in pre-verbal children* (pp. 192–200). Dordrecht, The Netherlands: Springer Netherlands.

Banks, M. S., & Dannemiller, J. L. (1987). Infant visual psychophysics. In P. Salapatek and L. Cohen (Eds.), *Developmental psychology series, volume 1: handbook of infant perception* (pp. 115–184). New York: Academic Press.

Banks, M. S., & Salapatek, P. (1981). Infant pattern vision: A new approach based on the contrast

sensitivity function. *Journal of Experimental Child Psychology, 31*(1), 1–45.

Banks, M. S., Stephens, B. R., & Dannemiller, J. L. (1982). A failure to observe negative preference in infant acuity testing. *Vision Research, 22*(8), 1025–1031.

British Society of Audiology. (2011). *Recommended procedure for pure tone air and bone conduction threshold audiometry with and without masking*. Reading, UK: British Society of Audiology.

Brown, A. M. (1990). Development of visual sensitivity to light and color vision in human infants: A critical review. *Vision Research, 30*(8), 1159–1188.

Brown, L. G. (1996). Additional rules for the transformed up-down method in psychophysics. *Perception & Psychophysics, 58*(6), 959–962.

Campbell, F. W., Carpenter, R. H. S., & Levinson, J. Z. (1969). Visibility of aperiodic patterns compared with that of sinusoidal gratings. *The Journal of Physiology, 204*(2), 283–298.

Carney, A. E. (1992). Bridging the gap between developmental psychoacoustics and pediatric audiology. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 333–349). Washington, DC: American Psychological Association, doi:10.1037/10119-007.

Cline, M. G. (1967). The perception of where a person is looking. *The American Journal of Psychology, 82*(4), 41–50, doi:10.2307/1420441.

Cone-Wesson, B. (2003). Pediatric audiology: A review of assessment methods for infants. *Audiological Medicine, 1*(3), 175–184.

Day, J., Green, R. J., Munro, K. J., Parry, G., Shaw, P., Wood, S. A., … Sutton, G. J. (2008). *Visual reinforcement audiometry testing of infants. A recommended test protocol (v. 2.0)*. Watford, UK: NHSP Clinical Group. Available at http://hearing.screening.nhs.uk/protocols_audioassess

Dobkins, K. R., Lia, B., & Teller, D. Y. (1997). Infant color vision: Temporal contrast sensitivity functions for chromatic (red/green) stimuli in 3-month-olds. *Vision Research, 37*(19), 2699–2716.

Dobson, V., & Teller, D. Y. (1978). Visual acuity in human infants: A review and comparison of behavioral and electrophysiological studies. *Vision Research, 18*(11), 1469–1483.

Fantz, R. L. (1958). Pattern vision in young infants. *The Psychological Record, 8,* 43–47.

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research, 38*(12), 1861–1881.

García-Pérez, M. A. (2001). Yes-no staircases with fixed step sizes: Psychometric properties and optimal setup. *Optometry & Vision Science, 78*(1), 56–64.

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *The Journal of the Acoustical Society of America, 87*(6), 2662–2674.

Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *The Journal of the Acoustical Society of America, 97*(6), 3749–3760.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Melbourne, FL: Krieger.

Gu, X., & Green, D. M. (1994). Further studies of a maximum-likelihood yes–no procedure. *The Journal of the Acoustical Society of America, 96*(1), 93–101.

Gwiazda, J., Wolfe, J. M., Brill, S., Mohindra, I., & Held, R. (1980). Quick assessment of preferential looking acuity in infants. *American Journal of Optometry and Physiological Optics, 57*(7), 420–427.

Harris, S. J., Hansen, R. M., & Fulton, A. B. (1984). Assessment of acuity in human infants using face and grating stimuli. *Investigative Ophthalmology & Visual Science, 25*(7), 782–786. [PubMed] [Article]

Held, R., Gwiazda, J., Brill, S., Mohindra, I., & Wolfe, J. (1979). Infant visual acuity is underestimated because near threshold gratings are not preferentially fixated. *Vision Research, 19*(12), 1377–1379.

Johnson, C. A., Chauhan, B. C., & Shapiro, L. R. (1992). Properties of staircase procedures for estimating thresholds in automated perimetry. *Investigative Ophthalmology & Visual Science, 33*(10), 2966–2974. [PubMed] [Article]

Jones, P. R., Kalwarowsky, S., Atkinson, J., Braddick, O. J., & Nardini, M. (2014). Automated measurement of resolution acuity in infants using remote eye-tracking. *Investigative Ophthalmology & Visual Science, 55*(12), 8102–8110. [PubMed] [Article]

Jones, P. R., Shub, D. E., Moore, D. R., & Amitay, S. (2015). The role of response bias in perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* E-pub ahead of print.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Attention, Perception, & Psychophysics, 49*(3), 227–229.

Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics, 63*(8), 1377–1388.

Kelly, D. H. (1970). Effects of sharp edges on the visibility of sinusoidal gratings. *Journal of the Optical Society of America, 60*(1), 98–102.

Kingdom, A. A. F., & Prins, N. (2009). *Psychophysics: A practical introduction*. London, England: Academic Press.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Attention, Perception, & Psychophysics, 63*(8), 1421–1455.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Attention, Perception, & Psychophysics, 63*(8), 1279–1292.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America, 49*(2), 467–477.

Lewis, T. L., Maurer, D., Chung, J. Y. Y., Holmes-Shannon, R., & Van Schaik, C. S. (2000). The development of symmetrical OKN in infants: Quantification based on OKN acuity for nasalward versus temporalward motion. *Vision Research, 40*(4), 445–453.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, D. L., Beiser, A. S., Warner, A. F., Pratt, E. M., Raye, K. N., & Lang, J. M. (1995). Monocular acuity norms for the Teller Acuity Cards between ages one month and four years. *Investigative Ophthalmology & Visual Science, 36*(3), 671–685. [PubMed] [Article]

McDonald, M. A., Dobson, V., Sebris, S. L., Baitch, L., Varner, D., & Teller, D. Y. (1985). The acuity card procedure: A rapid test of infant acuity. *Investigative Ophthalmology & Visual Science, 26*(8), 1158–1162. [PubMed] [Article]

McKee, S. P., Klein, S. A., & Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics, 37*(4), 286–298.

Mohn, G., & van Hof-van Duin, J. (1986). Rapid assessment of visual acuity in infants and children in a clinical setting, using acuity cards. In *Detection and measurement of visual impairment in pre-verbal children* (pp. 363–371). Dordrecht, the Netherlands: Springer Netherlands.

Moore, D. R., Ferguson, M. A., Halliday, L. F., & Riley, A. (2008). Frequency discrimination in children: Perception, learning and attention. *Hearing Research, 238*(1–2), 147–154.

Olsho, L. W., Koch, E. G., Halpin, C. F., & Carter, E. A. (1987). An observer-based psychoacoustic procedure for use with young infants. *Developmental Psychology, 23*(5), 627–640.

Pascalis, O., de Haan, M., & Nelson, C. A. (May 17, 2002). Is face processing species-specific during the first year of life? *Science, 296*, 1321–1323.

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision, 12*(6):25, 1–16, doi: 10.1167/12.6.25. [PubMed] [Article]

Rose, R. M., Teller, D. Y., & Rendleman, P. (1970). Statistical properties of staircase estimates. *Perception & Psychophysics, 8*(4), 199–204.

Saberi, K., & Green, D. M. (1996). Adaptive psychophysical procedures and imbalance in the psychometric function. *The Journal of the Acoustical Society of America, 100*(1), 528–536.

Salomão, S. R., & Ventura, D. F. (1995). Large sample population age norms for visual acuities obtained with Vistech-Teller Acuity Cards. *Investigative Ophthalmology & Visual Science, 36*(3), 657–670. [PubMed] [Article]

Swanson, W. H., & Birch, E. E. (1990). Infant spatiotemporal vision: Dependence of spatial contrast sensitivity on temporal frequency. *Vision Research, 30*(7), 1033–1048.

Swanson, W. H., & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics, 51*(5), 409–422.

Teller, D. Y. (1979). The forced-choice preferential looking procedure: A psychophysical technique for use with human infants. *Infant Behavior and Development, 2,* 135–153.

Teller, D. Y., Mar, C., & Preston, K. L. (1992). Statistical properties of 500-trial infant psychometric functions. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 211–227). Washington, DC: American Psychological Association, doi:10.1037/10119-008.

Teller, D. Y., Mayer, D. L., Makous, W. L., & Allen, J. L. (1982). Do preferential looking techniques underestimate infant visual acuity? *Vision Research, 22*(8), 1017–1024.

Teller, D. Y., McDonald, M. A., Preston, K., Sebris, S. L., & Dobson, V. (1986). Assessment of visual acuity in infants and children: The acuity card procedure. *Developmental Medicine & Child Neurology, 28*(6), 779–789.

Tobii Technology. (2012). *Accuracy and precision test report: Tobii X120 eye tracker*. Danderyd, Sweden: Tobii Technology AB.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*(17), 2503–2522.

Viemeister, N. F., & Schlauch, R. S. (1992). Issues in infant psychoacoustics. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp.

191–210). Washington, DC: American Psychological Association.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113–120.

Wattam-Bell, J. (1992). The development of maximum displacement limits for discrimination of motion direction in infancy. *Vision Research, 32*(4), 621–630.

Wattam-Bell, J. (1996). Visual motion processing in one-month-old infants: Preferential looking experiments. *Vision Research, 36*(11), 1671–1677.

Werner, L. A., & Bargones, J. Y. (1991). Sources of auditory masking in infants: Distraction effects. *Perception & Psychophysics, 50*(5), 405–412.

Werner, L. A., & Marean, G. C. (1991). Methods for estimating infant thresholds. *The Journal of the Acoustical Society of America, 90*(4), 1867–1875.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Attention, Perception, & Psychophysics, 63*(8), 1293–1313.

Wickens, T. D. (2002). *Elementary signal detection theory* (pp. 114–118). New York: Oxford University Press.

Widen, J. E., Folsom, R. C., Cone-Wesson, B., Carty, L., Dunnell, J. J., Koebsell, K., ... Norton, S. J. (2000). Identification of neonatal hearing impairment: Hearing status at 8 to 12 months corrected age using a visual reinforcement audiometry protocol. *Ear and Hearing, 21*(5), 471–487.